International Journal of
*Environmental Research
and Public Health*

MDPI

Article

The Comprehensive Machine Learning Analytics for Heart Failure

Chao-Yu Guo [1,2,*], Min-Yang Wu [1,2] and Hao-Min Cheng [1,2,3,4,*]

[1] Institute of Public Health, School of Medicine, National Yang-Ming University, Taipei 112, Taiwan; ben1423@hotmail.com.tw
[2] Institute of Public Health, School of Medicine, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[3] Center for Evidence-Based Medicine, Veteran General Hospital, Taipei 112, Taiwan
[4] Department of Medicine, School of Medicine, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
* Correspondence: cyguo@ym.edu.tw (C.-Y.G.); hmcheng@vghtpe.gov.tw (H.-M.C.)

Abstract: *Background*: Early detection of heart failure is the basis for better medical treatment and prognosis. Over the last decades, both prevalence and incidence rates of heart failure have increased worldwide, resulting in a significant global public health issue. However, an early diagnosis is not an easy task because symptoms of heart failure are usually non-specific. Therefore, this study aims to develop a risk prediction model for incident heart failure through a machine learning-based predictive model. Although African Americans have a higher risk of incident heart failure among all populations, few studies have developed a heart failure risk prediction model for African Americans. *Methods*: This research implemented the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression, support vector machine, random forest, and Extreme Gradient Boosting (XGBoost) to establish the Jackson Heart Study's predictive model. In the analysis of real data, missing data are problematic when building a predictive model. Here, we evaluate predictors' inclusion with various missing rates and different missing imputation strategies to discover the optimal analytics. *Results*: According to hundreds of models that we examined, the best predictive model was the XGBoost that included variables with a missing rate of less than 30 percent, and we imputed missing values by non-parametric random forest imputation. The optimal XGBoost machine demonstrated an Area Under Curve (AUC) of 0.8409 to predict heart failure for the Jackson Heart Study. *Conclusion*: This research identifies variations of diabetes medication as the most crucial risk factor for heart failure compared to the complete cases approach that failed to discover this phenomenon.

Keywords: heart failure; machine learning; prediction model; LASSO logistic regression; support vector machine; random forest; XGBoost

check for updates

Citation: Guo, C.-Y.; Wu, M.-Y.; Cheng, H.-M. The Comprehensive Machine Learning Analytics for Heart Failure. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4943. https://doi.org/10.3390/ijerph18094943

Academic Editor: Quynh C. Nguyen

Received: 26 March 2021
Accepted: 4 May 2021
Published: 6 May 2021

1. Introduction

The incidence and prevalence of heart failure have been increasing in recent years, and this phenomenon has become a serious global public health issue [1]. Research has estimated that more than 5.8 million people are suffering from heart failure in the United States, and this is increasing at a rate of at least 550,000 per year, while in the world, more than 23 million people are suffering from heart failure [2]. Other studies also pointed out that African Americans have a higher incidence of heart failure than other races [3,4], and even before the age of 50, the risk of heart failure of African Americans is 20 times that of whites [4].

Machine learning is equipped with algorithms that improve performance with experience [5]. A machine learns if its performance at a task improves with experience [6]. In constructing a model using machine learning techniques, the first step divides the entire data into two parts: the training data and the testing data. The training data refers to the dataset used to train the machine learning hyperparameters. The optimized model parameters are estimated based on the training datasets. The testing data is independent of the training data. The testing data will not participate in the training process but will only be used to evaluate the trained model. To avoid overfitting [7], K-Fold Cross Validation (CV) is a common strategy. The CV errors based on accuracy, Mean Square Error (MSE), and F1 Score are adopted [8].

Although it has been indicated that African Americans have a higher risk of incident heart failure among all populations [3], few studies have developed a risk prediction model for heart failure in African Americans. Therefore, this study utilized the Jackson Heart Study (JHS), the most prominent African American research database in the United States, and machine learning methods to construct comprehensive predictive models for heart failure. Our large-scale analysis structure aimed to discover missing and yet important information in such a big-data approach.

Random forest can provide feature importance through decision trees [9]. Breiman discussed the complete algorithm in 2001 [10], a type of ensemble method that collects multiple weak classifiers that produce a robust classifier [11]. The Classification and Regression Tree (CART) is a decision tree for predictive classification and continuous value [12] that adopts the binary division rule. Each time a division generates two branches, and the Gini classification method determines which branch is the best. Extreme Gradient Boosting (XGBoost) is a further extension from GBDT (Gradient Boosted Decision Tree) [13].

After missing data are correctly imputed, four methods, including (1) the Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression, (2) Support Vector Machine (SVM) [6,7], (3) random forest, and (4) Extreme Gradient Boosting (XGBoost), would be constructed for a predictive model for heart failure based on the African American population. The inclusion of the four strategies considers the essential aspects of the analysis concepts. LASSO logistic regression is the conventional statistical approach with a penalty term $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$ that avoids overfitting in regression models. The SVM is a machine learning tool dealing with classification problems. The SVM classifies subjects according to the separating hyperplane, which is defined as $\boldsymbol{\omega}^T \boldsymbol{x}_i + b = 0$, where $\boldsymbol{\omega}^T \boldsymbol{x}_i + b \geq 1, \forall y_i = 1$ and $\boldsymbol{\omega}^T \boldsymbol{x}_i + b \leq -1, \forall y_i = -1$. The last two methods are tree-based machine learning. However, the random forest and XGBoost are based on different mathematical optimizations. Random forest is an ensemble learning method that constructs a multitude of decision trees at training time. However, the XGBoost depends on gradient boosting trees, and the objection function is defined as $Obj^{(t)} = \sum_{j=1}^{T} \left[(G_j) w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T$, where $G_j = \sum_{i \in I_j} g_i$ is the sum of the first derivative of the loss function in the jth leaf and $H_j = \sum_{i \in I_j} h_i$ represents similar calculations for the second derivative.

As a result, we examined 112 scenarios, and Figure 1 displays the analysis plan. To date, this research had the most extensive models, with seven missing patterns for quality control, four missing imputation strategies, and four power machine learning models. We aimed to discover novel indications for heart failure from such a big scale of analysis models.

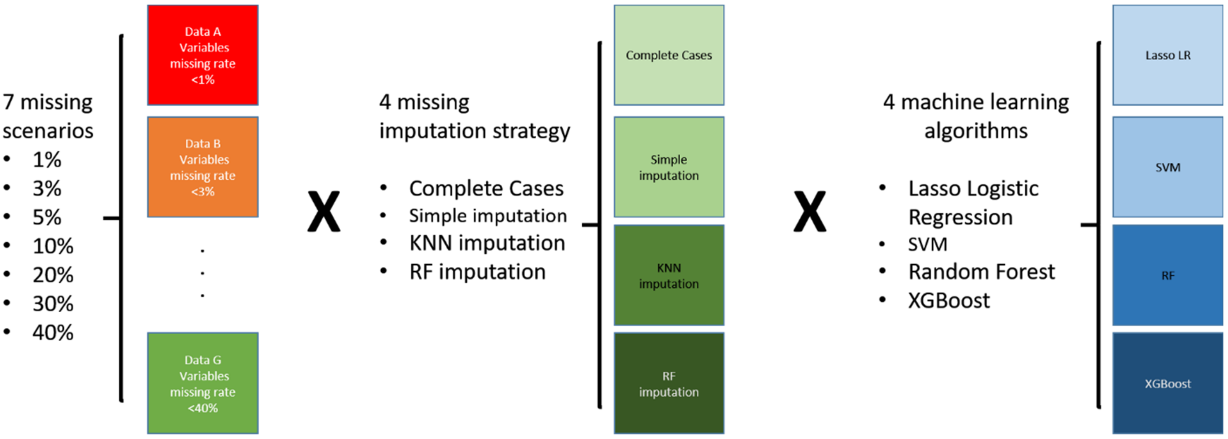
7 missing scenarios
1%
3%
5%
10%
20%
30%
40%
Data A Variables missing rate <1%
Data B Variables missing rate <3%
Data G Variables missing rate <40%
X
4 missing imputation strategy
Complete Cases
Simple imputation
KNN imputation
RF imputation
X
4 machine learning algorithms
Lasso Logistic Regression
SVM
Random Forest
XGBoost
Lasso LR
SVM
RF
XGBoost

Figure 1. Scenarios considered.

2. Materials and Methods

Following the previous work [14], the JHS initially included 3883 people. After excluding those who had experienced heart failures and those who were unsure whether they had heart failures or not at the baseline, 3327 people remained in the study. At the end of the study, 246 incident heart failure cases were identified. Table 1 shows the descriptive statistics. Subjects with heart failures accounted for 7.4% of the overall study sample.

Table 1. Descriptive statistics of the study population.

Baseline Characteristic	Total Population n = 3327	Non-HF n = 3081 (92.6%)	HF n = 246 (7.4%)
Age	54.96 (12.59)	54.24 (12.37)	63.91 (11.98)
BMI	31.82 (7.2)	31.72 (7.18)	33.02 (7.31)
Waist	100.83 (16.03)	100.4 (15.93)	106.9 (16.14)
High School Graduate	2761 (83.26%)	2607 (84.92%)	154 (62.60%)
Gender			
Male	1228 (36.91%)	1132 (36.74%)	96 (39.02%)
Female	2099 (63.09%)	1949 (63.26%)	150 (60.98%)
Current Smoker	406 (12.31%)	374 (12.25%)	32 (13.11%)
Hypertension (HTN)	1845 (55.47%)	1644 (53.38%)	201 (81.71%)
Diabetes Mellitus (DM)	710 (21.5%)	593 (19.39%)	117 (47.95%)

Regarding the missing data, variables in the database were classified according to the missing rates, which were <1%, <3%, <5%, <10%, <20%, <30%, and <40%. The missing rate was specific to the variable but not the proportion of missing data in the whole dataset. The column (<40%) means that this dataset contained all variables with missing rates less than 40%. Therefore, more variables were included in the analyses compared to other missing rates. Four imputation techniques dealt with the missing values, including (1) complete cases, (2) simple imputation, (3) the K Nearest-Neighbor (KNN) interpolation method [15], and (4) random forest imputation. More details of the imputation strategy and the statistical properties are available in a former article [16]. Besides, the random forest interpolation method is a machine learning-based imputation of the improved "MissForest" interpolation method [17].

After missing values were imputed, the next step was to generate dummy variables for the unordered category variables. This research adopted the Python language with "one-hot encoding" for categorical variables. We normalized continuous variables according to Min-Max normalization. The formula is $x' = \frac{x-\min(x)}{\max(x)-\min(x)}$. The Min-Max normalization limited the range of each variable in the dataset to [0, 1]. In this way, each variable had the same scale, which could improve the predictive performance by avoiding an extremely skewed distribution. Finally, we divided the dataset into the training set (70%) and the testing set (30%). Figure 2 demonstrates the flow chart of the entire data processing.

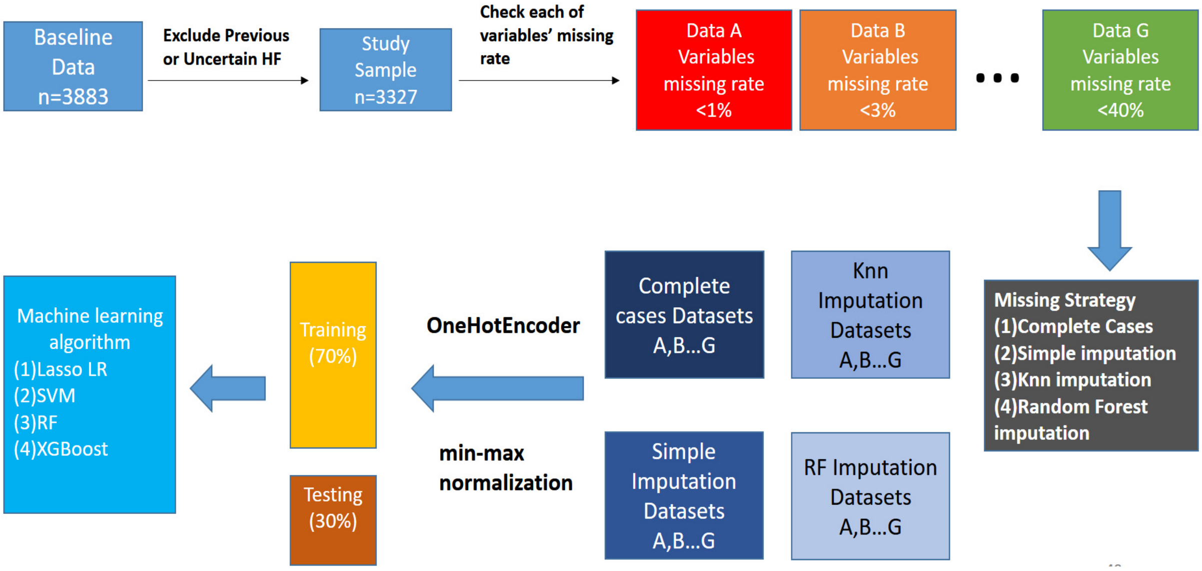
Baseline Data n=3883
Exclude Previous or Uncertain HF
Study Sample n=3327
Check each of variables' missing rate
Data A Variables missing rate <1%
Data B Variables missing rate <3%
• • •
Data G Variables missing rate <40%
Missing Strategy
(1)Complete Cases
(2)Simple imputation
(3)Knn imputation
(4)Random Forest imputation
Complete cases Datasets A,B…G
Knn Imputation Datasets A,B…G
Simple Imputation Datasets A,B…G
RF Imputation Datasets A,B…G
OneHotEncoder
min-max normalization
Training (70%)
Testing (30%)
Machine learning algorithm
(1)Lasso LR
(2)SVM
(3)RF
(4)XGBoost

Figure 2. Study flow chart.

Note that subjects with heart failures in this study population accounted for 7.4% of the overall sample. For machine learning models, this is an imbalance problem that introduces biases when building the model. We controlled the "class weight" in the LASSO logistic regression, SVM, and random forest to avoid potential biases. The "class weight" parameter in this paper is "balanced," and the formula is: $w_j = \frac{n}{kn_j}$, where "w_j" represents the weight of the jth category of the sample, "n" represents the sample size, "k" represents the number of categories, and "n_j" represents the sample size of the jth category. In the XGBoost, increasing the weight by adjusting the parameter "scale_pos_weight" could increase the predictive ability. The "scale_pos_weight" parameter finds the most appropriate value through K-fold cross-validation, and its range is between 1 and 10.

The logistic regression model is estimated by maximum likelihood [18], but we implemented the Least Absolute Shrinkage and Selection Operator (LASSO)-based model in this research [19]. The SVM [20] with various kernel functions were considered, including the linear kernel, polynomial kernel, and radial basis function kernel. Ten-fold cross-validation found the most suitable kernel function to establish the optimal SVM to predict heart failure.

This research's XGBoost model is based on 10-fold cross-validation to find the best parameters to establish a heart failure prediction model. The parameters and ranges are as follows: "scale_pos_weight": range 1~10. "n_estimators": initial training 90~150, final training 500~3500, "max_depth": range 2~10. "min_child_weight": range 1~10. "Gamma": range 0~5. "Subsample": range 0.3~1. "colsample_bytree": range 0.3~1. "reg_lambda": range 10^{-3}~10^{2}. "reg_alpha": range 10^{-3}~10^{2}. "learning_rate": range 0.01~0.1. As there are too many XGBoost parameters, it is time-consuming to train all the parameters simultaneously. Therefore, our procedure adopted training one set of two sets of parameters and gradually trained the XGBoost model (Figure 3).

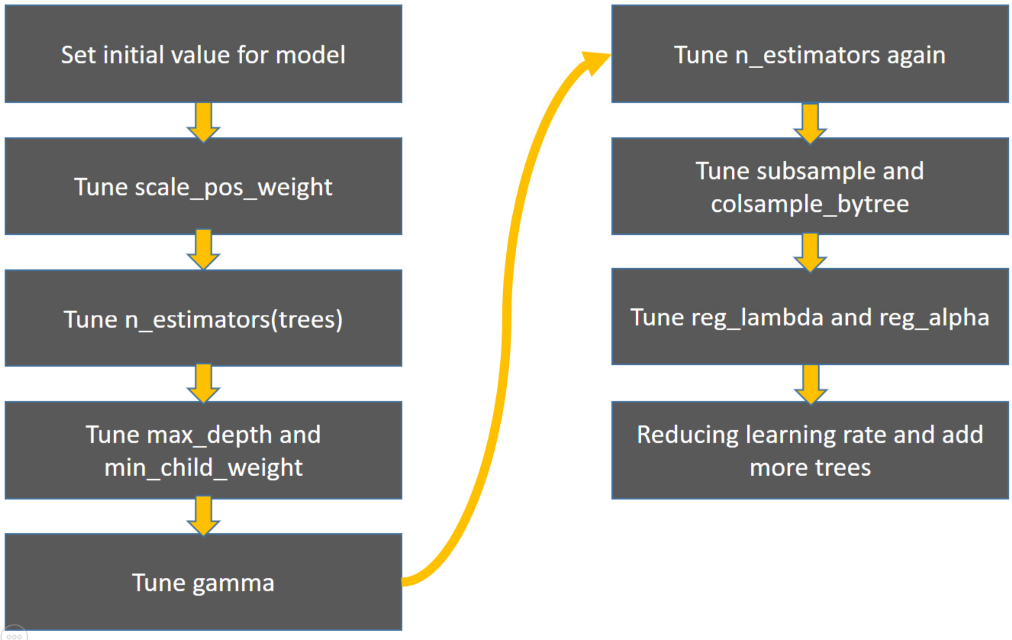
Set initial value for model
Tune scale_pos_weight
Tune n_estimators(trees)
Tune max_depth and min_child_weight
Tune gamma
Tune n_estimators again
Tune subsample and colsample_bytree
Tune reg_lambda and reg_alpha
Reducing learning rate and add more trees

Figure 3. The training process of XGBoost.

3. Results

If the analyses excluded missing data, Figure 4 shows the corresponding results, where the Y-axis is Area Under Curve (AUC), and the X-axis is the proportion of the missing value of the data. The predictive ability of XGBoost was better than the other three in most scenarios, and the best AUC was 0.808 when the missing rates of included variables were less than 3%. The best AUC of random forest was 0.8003 when the missing rates were less than 3%. The best AUC of SVM was 0.7892 when the missing rates were less than 3%. The best AUC of LASSO logistic regression was 0.7676 when the missing rates were less than 3%.

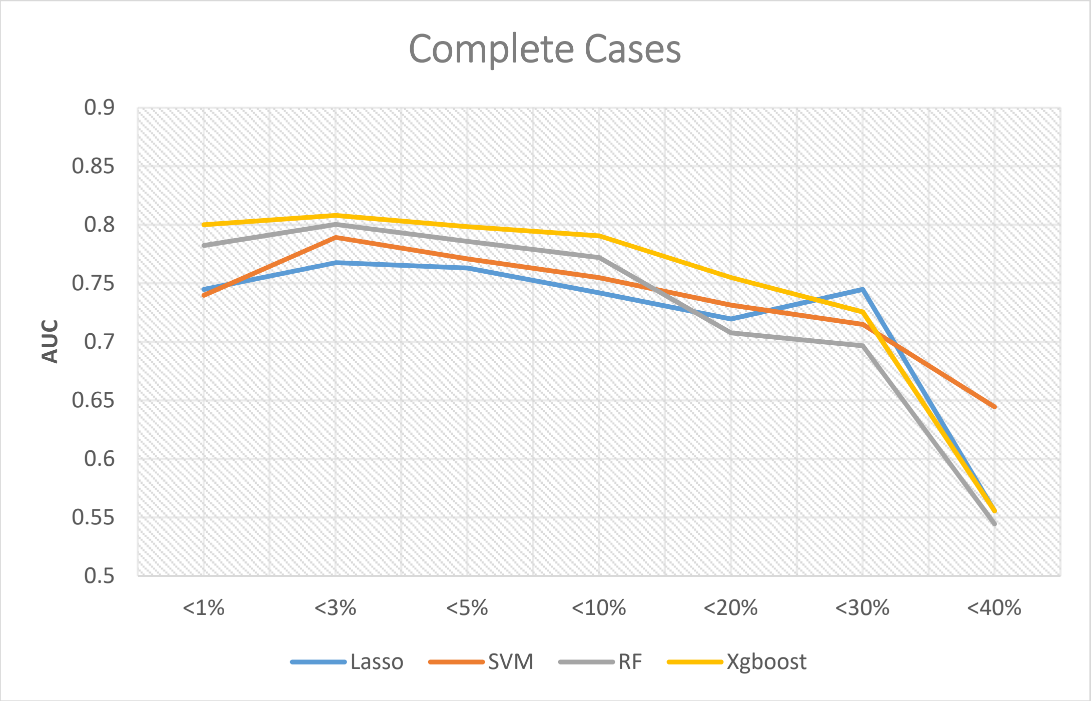

Figure 4. AUC for complete cases.

In summary, the best overall AUC performance occurs when the missing rates of included variables are less than 3%. When we included more variables in the analysis, we also excluded more subjects from the analysis due to missing values. As a result, the AUC decreased.

If the simple interpolation method is adopted, Figure 5 reveals the corresponding results. Similarly, XGBoost was the best performer, and the highest AUC was 0.8239 when the missing rates were less than 30%. The predictive abilities of random forest and LASSO logistic regression are very similar. The best AUC of random forest was 0.813, and LASSO logistic regression demonstrated a value of 0.814. Lastly, the AUC of SVM was 0.79 on average.

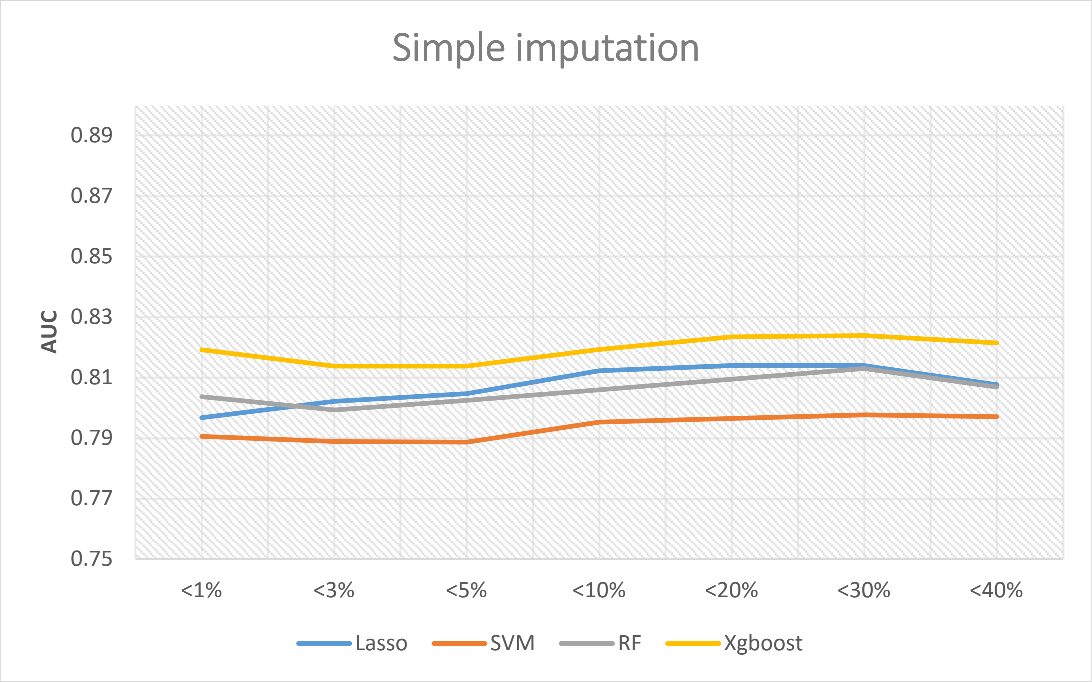

Figure 5. AUC with simple imputation.

In summary, the best overall AUC performance was when the proportion of missing values of the included variables was less than 30%. If we further included the variables whose missing value ratio was less than 40%, the AUC decreased. Therefore, if a follow-up researcher wants to use a simple interpolation method to establish a prediction model, the suggestion from the results of this paper is to include variables whose proportion of missing values is less than 30%. The prediction performance will be relatively good, of which the best performance was the XGBoost prediction model, which had an AUC of 0.8239.

If the KNN interpolation method is adopted, Figure 6 shows the corresponding results. Among all scenarios, XGBoost remained the best performer, and the highest AUC was 0.8369 if the missing rates were less than 30%. Random forest and LASSO logistic revealed similar performance. The best AUC of random forest was 08136 when the missing rates were less than 30%, and LASSO logistic regression was 0.8157 when the missing rates were less than 20%. Finally, the AUC of SVM was 0.79 if the missing rates were less than 30%. In summary, the best AUC occurs when the missing rates are less than 30%.

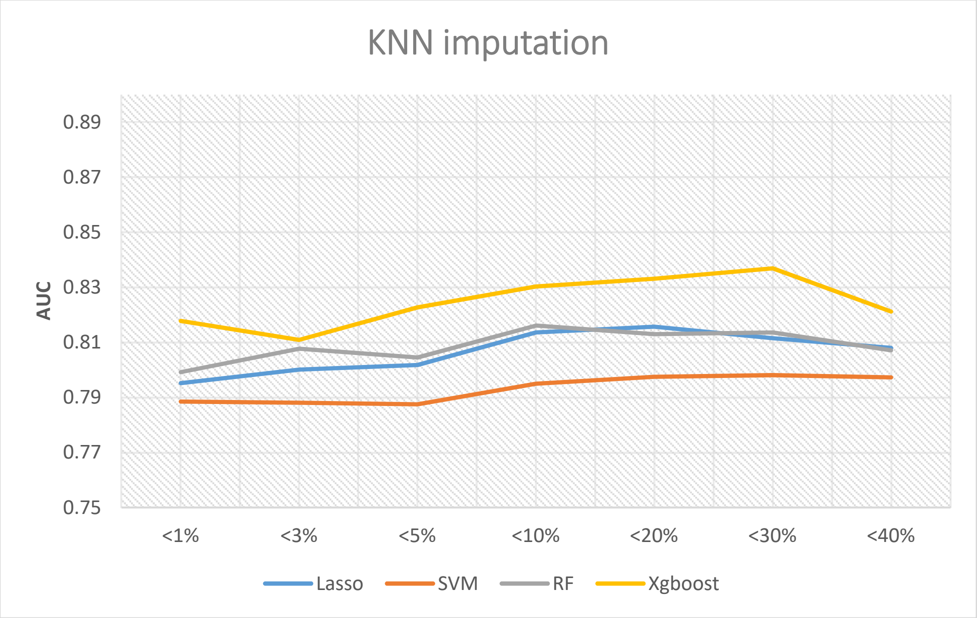

Figure 6. AUC with KNN imputation.

If the random forest interpolation method is adopted, Figure 7 shows the corresponding results. The XGBoost outperformed the other three methods in all analyses, and the best AUC was 0.8409 when the missing rates were less than 30%. Random forest and LASSO logic regression have similar predictive abilities. The best AUC of random forest was 0.8191 when the missing rates were less than 30%, while LASSO logic regression was 0.8166 when the missing rates were less than 20%. Lastly, the AUC of SVM was still about 0.79. In summary, the best AUC occurs when the missing rates are less than 30%.

According to these results, our analysis plan is an excellent example for future studies in various health outcomes. We conclude that even when variables have a missing rate as high as 30%, as long as we adopt the random forest imputation before the analysis the XGBoost provides the best predictive ability.

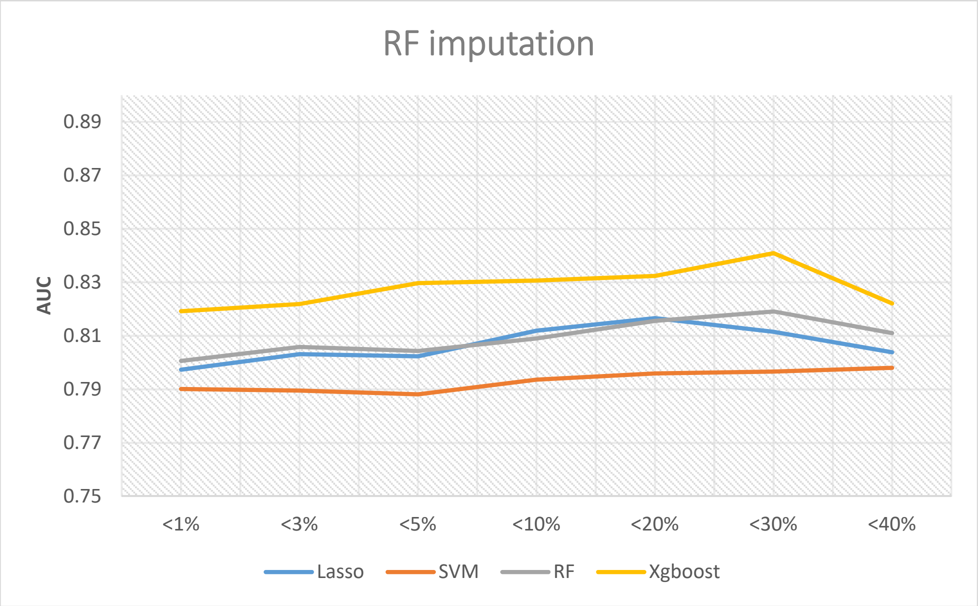

Figure 7. AUC with RF imputation.

4. Conclusions

In this research, the largest African American database in the United States, the JHS, was comprehensively examined by four different machine learning techniques to establish a heart failure predictive model. We also considered the impact of different missing rates and missing data imputation strategies on the predictive ability to discover the best analytic model. Besides, this work provides a series of guidelines for researchers interested in the JHS for future studies.

Based on these results, we conclude that (1) the XGBoost has the best performance in most situations to predict heart failure, especially when we implemented the random forest interpolation with the missing rates at the category of less than 30%. (2) The predictive performance using imputed data would be better than the models without imputations. Even the simplest imputation outperforms the complete cases scenario. (3) When building a predictive model, including too many variables (missing rates less than 40%) results in a worse prediction. The results revealed that in the analysis of complete cases, including variables with a missing rate of 3%, the predictive model performs better. Regarding the imputed data, it can tolerate up to 30% of the missing rates.

Finally, our work uses the feature importance function in random forest and XGBoost to discover which variables significantly contribute to predicting heart failures. Feature importance scores can provide insight and guidance in building the predictive model. We could use the importance scores to delete variables with the lowest scores and keep the variables with the highest scores.

Since the XGBoost performs better than random forest, we did not show the results of random forest. Tables 2–5 display details of the XGBoost. If we excluded missing data from the analysis, risk factors could not be identified consistently (Table 2). In contrast, if the analyses incorporated the imputed data, diabetes medication (DMmeds) was the most crucial feature suggested by the XGBoost results regardless of the imputation strategy. In Table 2, for data missing <5%, the feature importance of diabetes medication is 0.0492 compared to the age variable with 0.0290. Thus, diabetes medication is almost two times more important when building the predictive model (0.0492/0.029). DMmeds is a dichotomous variable. In this research, 15.25% of the population had a value of 1 with at least one diabetes medication. In total, 84.75% were without diabetes medication. Therefore, variations of diabetes medication were the most influential risk factor for heart failure.

Table 2. The selected parameters of the final HF prediction model for XGBoost using complete cases (feature importance is in the parenthesis).

<1%	**<3%**	**<5%**	**<10%**	**<20%**	**<30%**	**<40%**
age (0.0923)	age (0.0213)	DMmeds (0.0492)	DMmeds (0.0222)	dmMeds (0.0178)	CVDHx (0.0267)	frs_chdtenyrrisk (0.0507)
DMmeds (0.0647)	RepolarAntLat (0.0189)	age (0.0290)	Diabetes (0.0192)	MIHx (0.0174)	ascvd_tenyrrisk (0.0260)	ALDOSTERONE (0.0367)
Diabetes (0.0431)	DMmeds (0.0187)	BP3cat (0.0270)	CVDHx (0.0189)	EF (0.0170)	rrs_tenyrrisk (0.0220)	eGFRmdrd (0.0352)
eGFRckdepi (0.0315)	Diabetes (0.0180)	HTN (0.0267)	bpjnc7_3 (0.0187)	HbA1cIFCC (0.0150)	numbnessEver (0.0203)	occupation (0.0331)
MIecg (0.0305)	CVDHx (0.0174)	sbp (0.0237)	CHDHx (0.0176)	HbA1c (0.0148)	nutrition3cat (0.0194)	abi (0.0317)
RepolarAntLat (0.0297)	eGFRmdrd (0.0173)	eGFRckdepi (0.0233)	eGFRckdepi (0.0174)	strokeHx (0.0141)	FEV1PP (0.0193)	sbp (0.0297)
antiArythMeds (0.028)	eGFRckdepi (0.0173)	CVDHx (0.0188)	FPG (0.0170)	Diabetes (0.0141)	totchol (0.0186)	calBlkMeds (0.0292)
RepolarAnt (0.0272)	statinMeds (0.0168)	QTcFrid (0.0187)	age (0.0170)	visionLossEver (0.0140)	asthma (0.0180)	FVC (0.0289)

Table 2. *Cont.*

<1%	<3%	<5%	<10%	<20%	<30%	<40%
statinMeds	edu3cat	BPmeds	sbp	statinMeds	BPmeds	eGFRckdepi
(0.0268)	(0.0163)	(0.0184)	(0.0164)	(0.0129)	(0.0177)	(0.0253)
CVDHx	CardiacProcHx	waist	waist	frs_chdtenyrrisk	HTN	SCrCC
(0.0262)	(0.0162)	(0.01831)	(0.0160)	(0.0129)	(0.0177)	(0.0242)

Note: The above abbreviations are available in Table A1.

Table 3. The selected parameters of the final HF prediction model for XGBoost using simple imputations.

<1%	<3%	<5%	<10%	<20%	<30%	<40%
Diabetes	DMmeds	DMmeds	DMmeds	DMmeds	DMmeds	DMmeds
(0.0455)	(0.0346)	(0.0588)	(0.0233)	(0.0630)	(0.0210)	(0.0337)
DMmeds	age	Diabetes	age	age	age	DialysisEver
(0.0439)	(0.0273)	(0.0327)	(0.0221)	(0.0383)	(0.0196)	(0.0252)
age	CVDHx	DialysisEver	eGFRckdepi	Diabetes	Diabetes	CVDHx
(0.0408)	(0.0265)	(0.0243)	(0.0185)	(0.0297)	(0.0158)	(0.0207)
HTN	eGFRckdepi	MIAntLat	Diabetes	DialysisEver	BPmeds	age
(0.0336)	(0.0263)	(0.0191)	(0.0183)	(0.0190)	(0.0150)	(0.0151)
CVDHx	HTN	age	FEV1	ConductionDefect	CVDHx	Afib
(0.0277)	(0.0260)	(0.0189)	(0.0163)	(0.0183)	(0.0147)	(0.0142)
HSgrad	HSgrad	HSgrad	CVDHx	sex	age	MIHx
(0.0273)	(0.0246)	(0.0157)	(0.0158)	(0.0180)	(0.0145)	(0.0136)
BPmeds	Diabetes	Afib	FVC	occupation	ConductionDefect	eGFRckdepi
(0.0272)	(0.0237)	(0.0148)	(0.0156)	(0.0159)	(0.0141)	(0.0129)
eGFRckdepi	eGFRmdrd	edu3cat	CHDHx	MIant	FEV1	SystLVdia
(0.0238)	(0.0225)	(0.0138)	(0.0152)	(0.0143)	(0.0141)	(0.0128)
RepolarAntLat	MIHx	CVDHx	eGFRmdrd	CVDHx	eGFRckdepi	EF
(0.0229)	(0.0212)	(0.0135)	(0.0151)	(0.0129)	(0.0139)	(0.0117)
ecgHR	sbp	EF	HbA1cIFCC	idealHealthSMK	CHDHx	ConductionDefect
(0.0206)	(0.0191)	(0.0130)	(0.0148)	(0.0125)	(0.0126)	(0.0115)

Table 4. The selected parameters of the final HF prediction model for XGBoost using KNN imputations.

<1%	<3%	<5%	<10%	<20%	<30%	<40%
Diabetes	age	DMmeds	DMmeds	DMmeds	DMmeds	DMmeds
(0.0435)	(0.0320)	(0.0848)	(0.0261)	(0.0443)	(0.0145)	(0.0318)
DMmeds	DMmeds	age	age	age	age	DialysisEver
(0.0409)	(0.0266)	(0.0302)	(0.0177)	(0.0420)	(0.0142)	(0.0294)
age	MIHx	MIant	CVDHx	CVDHx	eGFRmdrd	age
(0.0386)	(0.0234)	(0.0245)	(0.0175)	(0.0321)	(0.0138)	(0.0234)
HTN	Diabetes	CVDHx	eGFRckdepi	EF	Diabetes	Diabetes
(0.0299)	(0.0213)	(0.0241)	(0.0173)	(0.0188)	(0.0130)	(0.0145)
CVDHx	CVDHx	Diabetes	Diabetes	eGFRckdepi	SCrCC	Afib
(0.0277)	(0.0209)	(0.0236)	(0.0167)	(0.0182)	(0.0122)	(0.0143)
BPmeds	eGFRckdepi	HSgrad	eGFRmdrd	FEV1	eGFRckdepi	CVDHx
(0.0274)	(0.0205)	(0.0206)	(0.0153)	(0.0179)	(0.0118)	(0.0128)
HSgrad	HSgrad	ConductionDefect	FEV1	ConductionDefect	statinMeds	eGFRckdepi
(0.0264)	(0.0176)	(0.0201)	(0.0153)	(0.0174)	(0.0115)	(0.0126)
eGFRckdepi	HTN	antiArythMedsSelf	CHDHx	MIHx	CVDHx	MIHx
(0.0222)	(0.0172)	(0.0152)	(0.0152)	(0.0172)	(0.0112)	(0.0120)
RepolarAntLat	antiArythMeds	CHDHx	edu3cat	MajorScarAnt	everSmoker	calBlkMeds
(0.0209)	(0.0171)	(0.1431)	(0.0142)	(0.0172)	(0.0111)	(0.0116)
ecgHR	eGFRmdrd	AntiArythMeds	DialysisEver	eGFRmdrd	rrs_tenyrrisk	FEV1
(0.0196)	(0.0164)	(0.1374)	(0.0139)	(0.0170)	(0.0108)	(0.0113)

Table 5. The selected parameters of the final HF prediction model for XGBoost using MissForest imputations.

<1%	<3%	<5%	<10%	<20%	<30%	<40%
Diabetes	antiArythMeds	dmMeds	DMmeds	DMmeds	DMmeds	DMmeds
(0.0455)	(0.0339)	(0.0340)	(0.0263)	(0.0443)	(0.0163)	(0.0290)
DMmeds	DMmeds	age	age	DialysisEver	Diabetes	ascvd_tenyrrisk
(0.0363)	(0.0332)	(0.0279)	(0.0251)	(0.0323)	(0.0161)	(0.0255)
age	age	Diabetes	Diabetes	MIAntLat	rrs_tenyrrisk	age
(0.0344)	(0.0301)	(0.0271)	(0.0234)	(0.0241)	(0.0148)	(0.0218)
HTN	eGFRckdepi	eGFRckdepi	CVDHx	Diabetes	age	eGFRckdepi
(0.0336)	(0.0241)	(0.0269)	(0.0218)	(0.0208)	(0.0132)	(0.0210)
CVDHx	HTN	CVDHx	CHDHx	age	ascvd_tenyrrisk	rrs_tenyrrisk
(0.0318)	(0.0232)	(0.0235)	(0.0194)	(0.0194)	(0.0130)	(0.0191)
HSgrad	SCrIDMS	eGFRmdrd	eGFRmdrd	Afib	MIant	frs_cvdtenyrrisk
(0.0274)	(0.0220)	(0.0212)	(0.0192)	(0.0184)	(0.0127)	(0.0179)
eGFRckdepi	MIHx	HSgrad	eGFRckdepi	calBlkMeds	eGFRckdepi	MIHx
(0.0243)	(0.0207)	(0.0198)	(0.0162)	(0.0154)	(0.0125)	(0.0162)
CHDHx	CVDHx	SCrIDMS	HSgrad	CVDHx	CVDHx	LEPTIN
(0.0241)	(0.0198)	(0.0187)	(0.0162)	(0.0152)	(0.0118)	(0.0147)
RepolarAntLat	eGFRmdrd	BPMeds	FEV1	eGFRckdepi	FEV1	calBlkMeds
(0.0238)	(0.0197)	(0.0170)	(0.0149)	(0.0149)	(0.0106)	(0.0135)
QTcBaz	Diabetes	HbA1c	SCrIDMS	EF	CHDHx	CardiacProcHx
(0.0221)	(0.0195)	(0.0158)	(0.0148)	(0.0140)	(0.0104)	(0.0127)

Furthermore, most of the findings revealed that heart failure was related to age, sex, diabetes, kidney function, and past heart disease history. Previous studies also pointed out that these variables are risk factors for heart failure [21–23], suggesting that our prediction models established by machine learning could discover crucial risk factors. In particular, a former study indicated significant predictors, including age, sex, body mass index, diabetes mellitus, systolic blood pressure, creatinine, serum albumin or total protein, LV hypertrophy, and coronary artery disease [24]. Compared to the simple AHEAD scoring system [25], the other AHEAD-U scoring system proposed by Chen et al. [26], and the most recent HANBAH score by Guo et al. [27], the most crucial feature importance variables determined by the XGBoost are similar to these variables used in the scoring system. However, variations of diabetes medications have never been reported previously. Therefore, the XGBoost discovered a novel influential risk factor and other similar factors that match previous studies.

In summary, the results suggest using variables with missing rates of less than 30% and adopting the Random Forest imputations when building a heart failure predictive model using the JHS database. Finally, we recommend that the XGBoost build the risk model, and the highest predictive accuracy is 0.8409. Most importantly, this research identifies variations of diabetes medication as the most crucial risk factor for heart failure, regardless of the missing data imputation strategy, since the variable of diabetes medications has the highest feater importance. In contrast, the commonly adopted approach that only uses complete cases failed to provide such novel findings.

The variable "diabetes medication" indicates that patients had diabetes, which is different from the conventional understanding of diabetic treatment [28]. The superficial understanding suggests that patients receiving diabetic treatment had a worse prognosis than those without treatment. It can be misunderstood that diabetes medication is the most crucial risk factor for heart failure. In this research, we interpret the variable "diabetes medication" as variations of diabetes medication since this variable has multiple levels, and we treated it as a dummy variable in the analysis.

In previous research [29,30], the authors used five methods to predict heart failure, including logistic regression, decision tree, random forest, simple Bayesian classifier, and support vector machine (SVM). The source of the data is from the Cleveland database. The number of samples included was 303, and the number of variables included was 14. For the data's missing values, the study used the K-Nearest Neighbor (KNN) interpolation method. In the final predictive results, the study adopted accuracy to conclude that the five methods have good prediction accuracy (87.36%, 93.19%, 89.14%, 87.27%, 92.30%), and the decision tree performed best.

Later, the random survival forest [31] built a predictive model. The database was Action to Control Cardiovascular Risk in Diabetes (ACCORD). There were 8756 samples included. After excluding variables with a missing value greater than 10%, the number of variables included was 109. Regarding the missing values, this research imputed missing data using random forest interpolation. The C-index evaluated the final predictive results, and the random survival forest yielded a value of 0.74.

Another study implemented three methods, including logistic regression, SVM, and Ada-Boost, to predict heart failure [32]. They came from the Geisinger Clinic Electronic Health Record system (Electronic Health Record). The sample size was 4489, and the number of variables was 179. For the missing values, this study used the missing indicator method. For model optimization and selection of parameters, this study adopted a 10-fold CV. The study used the Area Under the Curve (AUC) to compare the final predictive results. The Receiver Operating Characteristic (ROC) curve is an excellent diagnostic tool [33]. The Youden Index is also a popular choice [34] defined as $\operatorname{argmax}(\text{true positive rate } + \text{true negative rate} - 1,\ 1)$. The higher Area Under the Curve (AUC) [35] also indicates a better prediction [36]. Previous research suggested the interpretations of different cutoffs of AUC higher than 0.5 [37]. The logistic regression results were similar to the Ada-Boost with an AUC of 0.77, and the SVM was relatively unsatisfactory with an AUC of 0.62.

In future development, one could further discuss the following aspects. The problem of imbalanced data explains the weight of a small number of samples to be adjusted. There are many methods to deal with this issue, such as undersampling, which reduces multiple types of samples, or oversampling (SMOTE [38]) to generate more small samples, such that a ratio of normal and diseased samples of 1:1.

Other methods can build predictive models, such as the artificial neural network, naive Bayes, or other ensemble algorithms. These areas are essential topics in the future.

Author Contributions: C.-Y.G. proposed the study design; C.-Y.G. supervised statistical analyses; C.-Y.G. drafted the manuscript; M.-Y.W. carried out analysis and generated results; H.-M.C. secured funding support; H.-M.C. supervised the project and manuscript edits. All authors have read and agreed to the published version of the manuscript.

Funding: The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001), and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I, and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staff and participants of the JHS. Grants were received from the Ministry of Health and Welfare (MOHW104-TDU-B-211-113-003, MOHW106-TDU-B-211-113001) and an intramural grant from National Yang-Ming University (E107F-M01-0501). This work is particularly supported by “Yin Yen-Liang Foundation Development and Construction Plan” of the School of Medicine, National Yang-Ming University.

Institutional Review Board Statement: The study has been approved by the IRB of Taipei Veterans General Hospital.

Informed Consent Statement: Not applicable.

Data Availability Statement: This manuscript was prepared using the database of Jackson Heart Study with materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the Jackson Heart Study or the NHLBI. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Coding book for variables included in this research.

	Variable Name	Variable Types	Variable Description
1. Demographics			
	age	Continuous	Age in Years
	sex	Categorical	Participant Sex
	alc	Categorical	Alcohol drinking in the past 12 months (Y/N)
	alcw	Continuous	Average number of drinks per week
	currentSmoker	Categorical	Self-Reported Cigarette Smoking Status
	everSmoker	Categorical	Self-Reported History of Cigarette Smoking
2. Anthropometrics			
	weight	Continuous	Weight (kg)
	height	Continuous	Height (cm)
	BMI	Continuous	Body Mass Index (kg/m^2)
	waist	Continuous	Waist Circumference (cm)
	neck	Continuous	Neck Circumference (cm)
	bsa	Continuous	Calculated Body Surface Area (m^2)
	obesity3cat	Categorical	Ideal Health: BMI < 25 (Normal) Intermediate Health: 25 $\leq$ BMI < 30 (Overweight) Poor Health: BMI $\geq$ 30 (Obese)
3. Medications			
	medAcct	Categorical	Medication Accountability
	BPmedsSelf	Categorical	Self-Reported Blood Pressure Medication Status (Y/N)
	BPmeds	Categorical	Blood Pressure Medication Status (Y/N)
	DMmedsIns	Categorical	Diabetic Insulin Medication Status (Y/N)
	DMmedType	Categorical	Diabetes Medication Type
	dmMedsSelf	Categorical	Defined as Yes (Treated), if the participant reported being on diabetic
	DMmeds	Categorical	Diabetic Medication Status (Y/N)
	statinMedsSelf	Categorical	Defined as Yes (Treated), if the participant reported being on statin medication.
	statinMeds	Categorical	Statin Medication Status (Y/N)
	hrtMedsSelfEver	Categorical	Self Reported HRT Medication Status (Y/N)
	hrtMedsSelf	Categorical	Self Reported Current HRT Medication Status (Y/N)
	hrtMeds	Categorical	HRT Medication Status (Y/N)
	betaBlkMeds	Categorical	Beta Blocker Medication Status (Y/N)
	calBlkMeds	Categorical	Calcium Channel Blocker Medication Status (Y/N)
	diureticMeds	Categorical	Diuretic Medication Status (Y/N)
	antiArythMedsSelf	Categorical	Defined as Yes (Treated), if the participant reported being on antiarrhythmic medication.
	antiArythMeds	Categorical	Antiarrhythmic Medication Status (Y/N)
4. Hypertension			
	sbp	Continuous	Systolic Blood Pressure (mmHg)
	dbp	Continuous	Diastolic Blood Pressure (mmHg)
	BPjnc7	Categorical	JNC 7 BP Classification
	HTN	Categorical	Hypertension Status
	ABI	Continuous	Ankle Brachial Index

Table A1. *Cont.*

	Variable Name	Variable Types	Variable Description
5. Diabetes			
	FPG	Continuous	Fasting Plasma Glucose Level (mg/dL)
	FPG3cat	Categorical	Fasting Plasma Glucose Categorization
	HbA1c	Continuous	NGSP Hemoglobin HbA1c (%)
	HbA1c3cat	Categorical	NGSP Hemoglobin HbA1c (%) Categorization
	HbA1cIFCC	Continuous	IFCC Hemoglobin HbA1c in SI units (mmol/mol)
	HbA1cIFCC3cat	Categorical	IFCC Hemoglobin HbA1c in SI units (mmol/mol) Categorization
	fastingInsulin	Continuous	Fasting Insulin (Plasma IU/mL)
	HOMA-B	Continuous	HOMA-B
	HOMA-IR	Continuous	HOMA-IR
	Diabetes	Categorical	Diabetes Status (ADA 2010)
	diab3cat	Categorical	Diabetes Categorization
6. Lipids			
	ldl	Continuous	Fasting LDL Cholesterol Level (mg/dL)
	ldl5cat	Categorical	Fasting LDL Categorization
	hdl	Continuous	Fasting HDL Cholesterol Level (mg/dL)
	hdl3cat	Categorical	Fasting HDL Categorization
	trigs	Continuous	Fasting Triglyceride Level (mg/dL)
	trigs4cat	Categorical	Fasting Triglyceride Categorization
	totChol	Continuous	Fasting Total Cholesterol (mg/dL)
7. Biomarkers			
	hsCRP	Continuous	High Sensitivity C-Reactive Protein (Serum mg/dL)
	endothelin	Continuous	Endothelin-1 (Serum pg/mL)
	sCort	Continuous	Concentration of Cortisol Levels (Serum µg/dL)
	reninRIA	Continuous	Renin Activity RIA (Plasma ng/mL/hr)
	reninIRMA	Continuous	Renin Mass IRMA (Plasma pg/mL)
	aldosterone	Continuous	"Concentration of Aldosterone
	leptin	Continuous	(Serum ng/dL)"
	adiponectin	Continuous	Concentration of Leptin (Serum ng/mL)
8. Renal			
	SCrCC	Continuous	CC Calibrated Serum Creatinine (mg/dL)
	SCrIDMS	Continuous	IDMS Tracebale Serum Creatinine (mg/dL)
	eGFRmdrd	Continuous	eGFR MDRD
	eGFRckdepi	Continuous	eGFR CKD-Epi
	CreatinineU24hr	Continuous	24-hour urine creatinine (g/24hr)
	CreatinineUSpot	Continuous	Random spot urine creatinine (mg/dL)
	AlbuminUSpot	Continuous	Random spot urine albumin (mg/dL)
	AlbuminU24hr	Continuous	24-hour urine albumin (mg/24hr)
	DialysisEver	Categorical	Self-reported dialysis
	DialysisDuration	Continuous	Self-reported duration on dialysis (years)
	CKDHx	Categorical	Chronic Kidney Disease History
9. Respiratory			
	asthma	Categorical	Physician-Diagnosed Asthma
	maneuvers	Continuous	Successful Spirometry Maneuvers
	FVC	Continuous	Forced Vital Capacity (L)
	FEV1	Continuous	Forced Expiratory Volume in 1 s (L)
	FEV6	Continuous	Forced Expiratory Volume in 6 s (L)
	FEV1PP	Continuous	FEV1 % Predicted
	FVCPP	Continuous	FVC % Predicted

Table A1. *Cont.*

	Variable Name	Variable Types	Variable Description
10. Echocardiogram			
	LVMecho	Continuous	Left Ventricular Mass (g) from Echo
	LVMindex	Continuous	Left Ventricular Mass Indexed by Height(m)ˆ2.7
	LVH	Categorical	Left Ventricular Hypertrophy
	EF	Continuous	Ejection Fraction
	EF3cat	Categorical	Ejection Fraction Categorization
	DiastLVdia	Continuous	Diastolic LV Diameter (mm)
	SystLVdia	Continuous	Systolic LV Diameter (mm)
	FS	Categorical	Fractional Shortening
	RWT	Continuous	Relative Wall Thickness
11. Electrocardiogram			
	ConductionDefect	Categorical	Conduction Defect
	MajorScarAnt	Categorical	Anterior QnQs Major Scar
	MinorScarAnt	Categorical	Anterior QnQs Minor Scar
	RepolarAnt	Categorical	Anterior Repolarization Abnormality
	MIAnt	Categorical	Anterior ECG defined MI
	MajorScarPost	Categorical	Posterior QnQs Major Scar
	MinorScarPost	Categorical	Posterior QnQs Minor Scar
	RepolarPost	Categorical	Posterior Repolarization Abnormality
	MIPost	Categorical	Posterior ECG defined MI
	MajorScarAntLat	Categorical	Anterolateral QnQs Major Scar
	MinorScarAntLat	Categorical	Anterolateral QnQs Minor Scar
	RepolarAntLat	Categorical	Anterolateral Repolarization Abnormality
	MIAntLat	Categorical	Anterolateral ECG defined MI
	MIecg	Categorical	ECG determined MI
	ecgHR	Continuous	Heart Rate (bpm)
	Afib	Categorical	Atrial Fibrillation
	Aflutter	Categorical	Atrial Flutter
	QRS	Continuous	QRS Interval (msec)
	QT	Continuous	QT Interval (msec)
	QTcFram	Continuous	Framingham Corrected QT Interval (msec)
	QTcBaz	Continuous	Bazett Corrected QT Interval (msec)
	QTcHod	Continuous	Hodge Corrected QT Interval (msec)
	QTcFrid	Continuous	Fridericia Corrected QT Interval (msec)
	CV	Continuous	Cornell Voltage (microvolts)
	LVHcv	Categorical	Cornell Voltage Criteria
12. Stroke History			
	speechLossEver	Categorical	History of Speech Loss
	visionLossEver	Categorical	History of Sudden Loss of Vision
	doubleVisionEver	Categorical	History of Double Vision
	numbnessEver	Categorical	History of Numbness
	paralysisEver	Categorical	History of Paralysis
	dizzynessEver	Categorical	History of Dizziness
	strokeHx	Categorical	History of Stroke
13. CVD History			
	MIHx	Categorical	Self-Reported History of MI
	CardiacProcHx	Categorical	Self-Reported history of Cardiac Procedures
	CHDHx	Categorical	Coronary Heart Disease Status/History
	CarotidAngioHx	Categorical	Self-Reported history of Carotid Angioplasty
	CVDHx	Categorical	Cardiovascular Disease History
14. Healthcare Access			
	Insured	Categorical	Visit 1 Health Insurance Status

Table A1. *Cont.*

	Variable Name	Variable Types	Variable Description
15. Psychosocial			
	Income	Categorical	Income Status
	occupation	Categorical	Occupational Status
	edu3cat	Categorical	Education Attainment Categorization
	HSgrad	Categorical	High School Graduate
	dailyDiscr	Continuous	Everyday Discrimination Experiences
	lifetimeDiscrm	Continuous	Major Life Events Discrimination
	discrmBurden	Continuous	Discrimination Burden
	depression	Continuous	Total Depressive Symptoms Score
	weeklyStress	Continuous	Total Weekly Stress Score
	perceivedStress	Continuous	Total Global Stress Score
16. Life's Simple 7			
	SMK3cat	Categorical	AHA Smoking Categorization
	idealHealthSMK	Categorical	Indicator for Ideal Health via Smoking Status
	BMI3cat	Categorical	AHA BMI Categorization
	idealHealthBMI	Categorical	Indicator for Ideal Health via BMI
	PA3cat	Categorical	AHA Physical Activity Categorization
	idealHealthPA	Categorical	Indicator for Ideal Health via Physical Activity
	nutrition3cat	Categorical	AHA Nutrition Categorization
	idealHealthNutrition	Categorical	Indicator for Ideal Health via Nutrition
	totChol3cat	Categorical	AHA Total Cholesterol Categorization
	idealHealthChol	Categorical	Indicator for Ideal Health via Total Cholesterol
	BP3cat	Categorical	AHA BP Categorization
	idealHealthBP	Categorical	Indicator for Ideal Health via BP
	glucose3cat	Categorical	AHA Glucose Categorization
	idealHealthDM	Categorical	Indicator for Ideal Health via Glucose
17. Nutrition			
	vitaminD2	Continuous	25(OH) Vitamin D2 (ng/mL)
	vitaminD3	Continuous	25(OH) Vitamin D3 (ng/mL)
	vitaminD3epimer	Continuous	ep-25(OH) Vitamin D3 (ng/mL)
	darkgrnVeg	Continuous	Dark-green Vegetables
	eggs	Continuous	Eggs
	fish	Continuous	Fish
18. Physical Activity			
	sportIndex	Continuous	Sport Index
	hyIndex	Continuous	Home/Yard Index
	activeIndex	Continuous	Active Living Index
19. Risk Scores			
	frs_chdtenyrrisk	Continuous	Framingham Risk Score-Coronary Heart Disease
	frs_cvdtenyrrisk	Continuous	Framingham Risk Score-Cardiovascular Disease
	frs_atpiii_tenyrrisk	Continuous	Framingham Risk Score-Adult Treatment Panel (III)—Coronary Heart Disease
	rrs_tenryrisk	Continuous	Reynolds Risk Score
	ascvd_tenyrrisk	Continuous	American College of Cardiology—American Heart Association—Atherosclerotic Cardiovascular Disease

References

1. Conrad, N.; Judge, A.; Tran, J.; Mohseni, H.; Hedgecott, D.; Crespillo, A.P.; Allison, M.; Hemingway, H.; Cleland, J.G.; McMurray, J.J.V.; et al. Temporal trends and patterns in heart failure incidence: A population-based study of 4 million individuals. *Lancet* **2018**, *391*, 572–580. [CrossRef]
2. Roger, V.L. Epidemiology of heart failure. *Circ. Res.* **2013**, *113*, 646–659. [CrossRef] [PubMed]
3. Bahrami, H.; Kronmal, R.; Bluemke, D.A.; Olson, J.; Shea, S.; Liu, K.; Burke, G.L.; Lima, J.A. Differences in the incidence of congestive heart failure by ethnicity: The multi-ethnic study of atherosclerosis. *Arch. Intern. Med.* **2008**, *168*, 2138–2145. [CrossRef] [PubMed]
4. Bibbins-Domingo, K.; Pletcher, M.J.; Lin, F.; Vittinghoff, E.; Gardin, J.M.; Arynchyn, A.; Lewis, C.E.; Williams, O.D.; Hulley, S.B. Racial differences in incident heart failure among young adults. *N. Engl. J. Med.* **2009**, *360*, 1179–1190. [CrossRef]

5. Langley, P. *Elements of Machine Learning*; Morgan Kaufmann: Burlington, MA, USA, 1996.
6. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
7. Leinweber, D.J. Stupid data miner tricks: Overfitting the S&P 500. *J. Invest.* **2007**, *16*, 15–22.
8. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 569–575. [CrossRef] [PubMed]
9. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
10. Breiman, L.J. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
11. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [CrossRef]
12. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC press: Boca Raton, FL, USA, 1984.
13. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
14. Casanova, R.; Saldana, S.; Simpson, S.L.; Lacy, M.E.; Subauste, A.R.; Blackshear, C.; Wagenknecht, L.; Bertoni, A.G. Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. *PLoS ONE* **2016**, *11*, e0163942. [CrossRef] [PubMed]
15. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
16. Jonsson, P.; Wohlin, C. An evaluation of k-nearest neighbour imputation using likert data. In Proceedings of the 10th International Symposium on Software Metrics, Chicago, IL, USA, 22 November 2004; pp. 108–118.
17. Stekhoven, D.J.; Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef] [PubMed]
18. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2000.
19. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
20. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT'92), Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
21. Bauters, C.; Lamblin, N.; Mc Fadden, E.P.; Van Belle, E.; Millaire, A.; De Groote, P. Influence of diabetes mellitus on heart failure risk and outcome. *Cardiovasc. Diabetol.* **2003**, *2*, 1. [CrossRef]
22. Damman, K.; Van Veldhuisen, D.J.; Navis, G.; Vaidya, V.S.; Smilde, T.D.; Westenbrink, B.D.; Bonventre, J.V.; Voors, A.A.; Hillege, H.L. Tubular damage in chronic systolic heart failure is associated with reduced survival independent of glomerular filtration rate. *Heart* **2010**, *96*, 1297–1302. [CrossRef] [PubMed]
23. Metra, M.; Cotter, G.; Gheorghiade, M.; Dei Cas, L.; Voors, A.A. The role of the kidney in heart failure. *Eur. Heart J.* **2012**, *33*, 2135–2142. [CrossRef] [PubMed]
24. Echouffo-Tcheugui, J.B.; Greene, S.J.; Papadimitriou, L.; Zannad, F.; Yancy, C.W.; Gheorghiade, M.; Butler, J. Population risk prediction models for incident heart failure: A systematic review. *Circ. Heart Fail.* **2015**, *8*, 438–447. [CrossRef]
25. Spinar, J.; Jarkovsky, J.; Spinarova, L.; Mebazaa, A.; Gayat, E.; Vitovec, J.; Linhart, A.; Widimsky, P.; Miklik, R.; Zeman, K.; et al. AHEAD score–Long-term risk classification in acute heart failure. *Int. J. Cardiol.* **2016**, *202*, 21–26. [CrossRef]
26. Chen, Y.-J.; Sung, S.-H.; Cheng, H.-M.; Huang, W.-M.; Wu, C.-L.; Huang, C.-J.; Hsu, P.-F.; Yeh, J.-S.; Guo, C.-Y.; Yu, W.-C.; et al. Performance of AHEAD Score in an Asian Cohort of Acute Heart Failure With Either Preserved or Reduced Left Ventricular Systolic Function. *J. Am. Heart Assoc.* **2017**, *6*, e004297. [CrossRef]
27. Guo, C.Y.; Chan, C.H.; Chou, Y.C.; Sung, S.H.; Cheng, H.M. A Statistical Predictive Model Consistent Within a 5-Year Follow-up Period for Patients with Acute Heart Failure. *J. Chin. Med Assoc.* **2020**, in press. [CrossRef]
28. Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
29. Swedberg, K. Heart failure subtypes: Pathophysiology and definitions. *Diabetes Res. Clin. Pract.* **2021**, *13*, 108815. [CrossRef]
30. Gvozdanović, Z.; Farčić, N.; Šimić, H.; Buljanović, V.; Gvozdanović, L.; Katalinić, S.; Pačarić, S.; Gvozdanović, D.; Dujmić, Ž.; Miškić, B.; et al. The Impact of Education, COVID-19 and Risk Factors on the Quality of Life in Patients with Type 2 Diabetes. *Int. J. Environ. Res. Public Health* **2021**, *27*, 2332, PMCID:PMC7956830. [CrossRef] [PubMed]
31. Segar, M.W.; Vaduganathan, M.; Patel, K.V.; McGuire, D.K.; Butler, J.; Fonarow, G.C.; Basit, M.; Kannan, V.; Grodin, J.L.; Everett, B. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care* **2019**, *42*, 2298–2306. [CrossRef]
32. Wu, J.; Roy, J.; Stewart, W.F. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Med Care* **2010**, *48* (Suppl. 6), S106–S113. [CrossRef] [PubMed]
33. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
34. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35. [CrossRef]
35. Rosner, B. *Fundamentals of Biostatistics*; Nelson Education: Boston, MA, USA, 2015.
36. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]

37. Mandrekar, J.N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **2010**, *5*, 1315–1316. [CrossRef] [PubMed]
38. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]